Instrumentalist Theories of Moral Responsibility

Manuel Vargas
University of California, San Diego

Forthcoming in Nelkin and Pereboom, eds., *The Oxford Handbook on Moral Responsibility*

ABSRACT: In both the historical and contemporary literature on moral responsibility, there are accounts of responsibility that appeal to instrumentalist considerations in accounting for, variously, the nature, structure, and justification of moral responsibility. On the face of it, instrumentalist approaches can seem ill-suited for delivering an adequate theory of responsibility. For example, if one thinks that the hallmark of moral responsibility is its retrospective or "backward-looking" focus, that it involves some notion of desert, or that it somehow invokes or justifies retributive attitudes, instrumentalist theories of responsibility can seem to be non-starters. Even so, instrumentalist (or "consequentialist") approaches to responsibility have a storied pedigree within analytic philosophy. In recent years the approach has enjoyed renewed attention and rehabilitation.

This chapter provides an overview of instrumentalist theories of responsibility, including their history, recent developments, and ongoing disputes.

Keywords: consequentialism about responsibility; instrumentalism; desert; moral responsibility

Instrumentalist Theories of Moral Responsibility

Manuel Vargas
University of California, San Diego

1. Instrumentalism vs. consequentialism

There is no single, substantive commitment shared by instrumentalist approaches to responsibility. Instead, instrumentalist theories are best understood as a family of views that employ instrumentalist or consequentialist considerations in a variety of often distinct theoretical roles. For example, some accounts have maintained that individual instances of *holding* responsible are simply forward-looking attempts to modify behavior (Nowell-Smith 1948; Smart 1961); others have argued that what responsibility *is* just is a general practice of pro-social influence (Schlick 1939; Dennett 1984); still others have argued that whatever the right account of the nature of responsibility and the practice holding people responsible, the *justification* of responsibility is properly understood in instrumentalist terms (Vargas 2008, 2013a; McGeer 2013, 2015; Jefferson 2019).

It is common to refer to this variegated family of theories as "consequentialist." The description is informative and misleading in equal measure. In the 20th century, most instrumentalist theories of responsibility were held by people who were also consequentialists about normative ethics and practical reasoning, and who tended to see the development of an instrumentalist account of responsibility as an application of a broader commitment to consequentialism. Thus, *consequentialist* was an apt descriptor of those views.

Over the past few decades, instrumentalist theories of responsibility have come to be held by people who are either not committed to consequentialism, or who see the commitment to instrumentalism about responsibility as at least in principle separable from a commitment to consequentialism about normative ethics. To characterize these (at least in principle) independently motivated instrumentalist accounts as consequentialist can misleadingly suggest that these accounts necessarily involve a commitment to consequentialism about practical reason or normative ethics.

Although the label *consequentialist* is faithful to the tradition, for the sake of clarity I will follow Jefferson (2019) in using *instrumentalist* to refer to the general class of theories and *consequentialist* to refer to instrumentalist theories that are in some notable way connected to consequentialism about practical reason or normative ethics.[1]

2. A provisional sketch

As noted above, perhaps the most important traditional motivation for adopting an instrumentalist theory of responsibility was the (formerly) widespread acceptance of consequentialism in normative ethics. However, for those unpersuaded by consequentialism, this was no appeal at all. Still, there are reasons connected to debates about free will that seem to

[1] Although it raises interesting questions about the taxonomy of theories of responsibility, this chapter takes no stand on the relationship of instrumentalist accounts to recent functionalist accounts (e.g., in Shoemaker and Vargas forthcoming; and arguably McKenna 2012 and Smith 2013).

have made instrumentalist approaches appealing to some philosophers. This section offers a brief overview of the current state of play, and its connection to issues in debates about free will.

The nature of free will is a famously vexed matter, and the most traditional way of characterizing the partisans involves distinguishing between those who think free will is incompatible with determinism (incompatibilists) and those who do not (compatibilists).[2] Many incompatibilists have thought that free will's requirements are relatively robust, metaphysically speaking. For example, free will has been thought to involve "contra-causal" powers, causa sui, indeterministic causation, agent causation, or high-level emergent causes. Such esoterica—"panicky metaphysics," in Peter Strawson's (1962) memorable phrase—has struck many observers as either undermotivated, ad hoc, or at odds with independently plausible pictures of the world. The general upshot is that responsibility skepticism can seem an appealing view for those who think that these putatively unattainable forms of agency characterize the correct metaphysics of freedom (G. Strawson 1994, Pereboom 2001, Caruso 2012).

Enter instrumentalist theories. The metaphysics of agency required to make sense of instrumentalist theories tend to be had comparatively unproblematic. On the simplest versions (ones on which some observers will add, "by which you mean simplistic")—it is enough that agents be influenceable by praise and blame. Influenceability is a low metaphysical bar to cross, but it gets many of the cases right: where praise and blame have no hope of altering the dispositions of agents, praise and blame can seem to lose their luster. The appeal of instrumentalist accounts is that they involve a relatively uncontroversial metaphysics of agency while appearing to preserve the practice more or less as we find it, while also explaining why the practice is normatively appealing (about which more below).

The putative appeal of the instrumentalist's metaphysics of agency can strike some incompatibilists as an instance of philosophical sour grapes: the instrumentalist denies that we ever wanted what (to the incompatibilist) we manifestly *do* want of a metaphysics of agency. The dialectical issues here are tricky, and familiar to wider debates between compatibilists and incompatibilist. Partly, the issue turns on what one thinks of as the task of philosophy, whether it is methodologically conservative—e.g., a matter of descriptive metaphysics (Strawson 1962), or vindicatory, showing how our ordinary beliefs might be true (Nozick 1981)—or instead a matter of potential revisionary or even eliminativist possibilities. These reforming adjustments, whether revisionist or eliminativist can in turn be structured by different considerations about the relative importance of our practices versus our conceptual conjectures in fixing meaning and reference (Double 1996; Vargas 2011, 2013b; Nichols 2015; Caruso 2015; McCormick 2016; in the broader literature, see Rawls 1971; Daniels 1979; Jackson 1998).

Methodological concerns aside, simple instrumentalist approaches are vulnerable to challenges about their extensional adequacy. In actual practice, we find that some agents may be influenceable by blame and condemnation without being responsible (for examples, pets and

[2] This is a traditional way of construing the terrain, but in the current literature "determinism" is usually meant as a metaphysical thesis, but it is sometimes treated as a placeholder for a range of freedom-threatening things, which can include, in various formulations, physicalism, causal explanations in general, the causal closure of the physical, mechanism, naturalism, and determinism of more and less particular forms (such as psychological, neurological, biological, and physical determinism).

infants). Classical consequentialist theories seldom had a clear way of addressing the challenge. Contemporary instrumentalist theories of responsibility have tended to make use of explicit conditions on who is properly subject to praise and blame. Important candidates have included reasons-responsive or socially-sensitive reactivity to reasons (Vargas 2008, 2013a; McGeer 2014; Fricker 2016; Jefferson 2019). Like traditional consequentialist accounts, these newer accounts rely on powers that few doubt are compatible with determinism, but they provide more resources for getting the cases right: animals might be influenceable but not suitably reasons-responsive, and thus not blameworthy. Different strategies are sometimes available for capturing cases where agents seem blameworthy but not influenceable. For "two-tiered" or indirect versions of instrumentalism, the idea is that blameworthiness is a status in a justified practice, where the practice is instrumentally justified but the statuses within it make no appeal to instrumental considerations (Vargas 2013a). On "one-tiered" or direct approaches, there is the possibility of appealing to various indirect effects, and the instrumentalist may also insist that the phenomenology of blame may come apart from its instrumentally best function (Miller 2014a; McGeer 2019).

A different family of motivations for instrumental theories of responsibility turns on instrumentalism's relatively ready answer to doubts about the rational or normative grounds for responsibility practices. On these approaches, responsibility practices have some claim on us to the extent to which they generate morally desirable outcomes. Philosophers can and have readily disputed the appeal of, for example, retributive practices, or practices of blame that depend on metaphysically dubious notions of desert. However, the appeal of encouraging good actions, dispositions, or forms of agency and of discouraging bad instances of the same seems straightforward. Instrumentalists try to extend this appeal to a more general account of why we have reason to encourage and sustain practices of moralized praising and blaming. Whether such extensions must necessarily fail—whether they confuse the efficacy of holding responsible for the truth of one's being responsible, for example—is a matter we will return to, below.

If an antecedent commitment to consequentialism was once the main basis for traditional instrumentalist theories, it is one of the many ironies in the rehabilitation of this family of views that they tend to make relatively thin demands on one's theory of normative ethics. The reasoning is as follows. Given that instrumentalist considerations are allowed on most theories of normative ethics—although they may be subject to side-constraints or principled restrictions—a theory of moral responsibility that employs those considerations is not invoking elements that are automatically excluded from many theories of normative ethics. Instrumentalist approaches can therefore promise a relatively "thin" or non-partisan way of accounting for moral responsibility that is, at least in principle, compatible with a range of normative ethical theories (Arneson 2003; Vargas 2013a). If one goes on to accept some contractualist or deontological account of normative ethics, then it is conceivable that one's theory of normative ethics may provide additional side-constraints that, for example, prohibit scapegoating the non-responsible. In contrast, if one accepts a normative ethics that is purely consequentialist, then there is no basis for objecting that an instrumentalist theory of responsibility introduces special worries about scapegoating the non-responsible.[3]

[3] To be sure, important versions of contemporary instrumentalist accounts are explicitly wedded to a broader consequentialism about normative ethics (e.g., McGeer 2015; Miller 2014a, 2014b). The point here is about the resources for instrumentalist accounts that aren't wedded to consequentialism.

Independence from consequentialism about practical reasoning and normative ethics brings with it a number of advantages. Even so, instrumentalist approaches to responsibility that are not grounded in an antecedent commitment to consequentialism raise new puzzles. First, holding that there is a potentially autonomous normative domain of responsibility that is discreet from questions about normative ethics invites questions about the relationship of those normative domains, and whether one trumps or constrains the other. Second, the introduction of a free-standing consequentialist account does not obviously block the possibility that one could develop a distinct account of the normative foundations of responsibility grounded in some other view of normative ethics. How these potentially competing and non-overlapping accounts might interact raises difficult questions for the non-consequentialist instrumentalist about responsibility.

To better understand the current state of play—and to better see what progress remains to be made—it will be helpful to first canvass some of the historical origins the approach.

3. Classical consequentialist theories

If one thinks of Strawson's "Freedom and Resentment" as ushering in the contemporary era of work on moral responsibility, it is perhaps fair to say that for most of that post-Strawsonian history, consequentialist theories of responsibility have been treated as bogeymen—something seldom clearly seen, but universally regarded as bad. One source of this reputation was rooted in a prominent reading of Strawson's "Freedom and Resentment." Strawson's so-called optimist was committed to a "one-eyed consequentialism" that understood responsibility as a matter of the efficacy of moral influence. Consequentialists missed our concern for the quality of will evinced in action, and failed to recognize the central role of the reactive attitudes in shaping our practices.

Given the formidable grip that consequentialism exercised over the imagination of some philosophers in the first half of 20th century analytic philosophy, it is tempting to portray the main theoretical options in the theory of responsibility as a choice between some or another version consequentialism about responsibility, on the one hand, and metaphysical libertarianism on the other (Cf. Moore 1903; Schlick 1939; Nowell-Smith 1948). This is accurate enough, but it also overplays the role of consequentialism in that period. The bulk of the pre-Strawsonian literature tended to focus less on the normative foundations of responsibility and more on the meaning and metaphysics of ability talk. Consequentialist considerations were oftentimes in the background. However, in many of the discussions of that era, the philosophical heavy-lifting of compatibilism was performed by invocations of the supposed ordinary meanings of 'can', 'action', and 'volition' and by appeal to various broadly Humean pictures of freedom and responsibility (for examples, see Hook 1958, Pears 1963; and Dworkin 1970).

There were exceptions where consequentialism, rather than Humeanism about freedom and action, figured more prominently in the explanatory approach. For example, Schlick (1939) claimed that clarity about responsibility can be gotten by looking to the actual role our concept plays in ordinary life. On his account, our interest in responsibility and punishment is, at bottom, about shaping dispositions. We blame and punish to shape motives in ways that are designed to prevent or call forth certain acts. Thus, we don't punish agents for whom those punishments have no plausible influence on that agent's motives (1939, 151-4). Where some saw the consequentialist motive as built into the explicit aspirations or intentions of a blamer or punisher (e.g., Nowell-Smith 1948), Schlick's analysis suggests that the consequentialism is at the level of something like the functional role of responsibility attributions. For Schlick, we can understand the nature of responsibility by the role of the concept in our practices. Individual cases of blaming and punishing, it seems, might not be undertaken with a teleological end. Indeed, Schlick gives an error theory for how that might be so (154-8).

Other theories held that individual tokens of blaming were always aiming at character modification. A striking version of this is Nowell-Smith's (1948) view. Nowell-Smith was the famous "optimist" of Strawson's "Freedom and Resentment."[4] Echoing some of the ideas in Schlick, Nowell-Smith holds that the crucial idea for responsibility is that "the class of actions generally agreed to be voluntary coincides roughly with the class of actions that are caused by characteristics that can be strengthened or inhibited by praise and blame" (1948, 56). And, like Schlick, he quickly moves from the loose fit between blame and character modification to talk of punishment. On his view, this picture of the domain on the voluntary and its susceptibility to influence entails a utilitarian theory of punishment. However, unlike Schlick, Nowell-Smith holds that the intention of the punisher (and, it seems, the blamer) in so punishing *just is* to bring about particular effects. Failures to achieve those results don't show that punishment lacks instrumental intent, but rather, those failures simply speak to the "lack of skill in the practitioner" (56).

A third version of classical consequentialism about responsibility worth highlighting is J.J.C. Smart's "Free Will, Praise, and Blame" (1961). Given that its publication date is close to the publication of "Freedom and Resentment" it seems unlikely that Smart was one of Strawson's targets, although his account is vulnerable to some of Strawson's objections. Even so, Smart's proposal has continued to enjoy some visibility in part because of its distinctive character and its curious ambiguity about whether it is a theory of moral responsibility at all.

Akin to Schlick, Smart presents his view as getting at the conceptual role of responsibility practices. However, unlike Schlick, Smart finds that ordinary practices are a confused mix of tenable and untenable features. In particular, moral blame involves a picture of desert-entailing judgments that Smart regards as metaphysically confused. So, he advocates a relatively radical recasting of responsibility practices, jettisoning desert-entailing judgments of blame for a notion of "dispraise" that amounts to a kind of "grading" of the moral quality of the action.

Smart's account can be championed as both a compatibilist and incompatibilist account of responsibility.[5] Which way we read it partly depends on what we think fixes the target for

[4] Strawson's presentation of his optimist reads as though it was supposed to be about a general class of views, rather than any particular figure. And, indeed, in that time period the kind of view he was gesturing at would have been recognizable to his audience. However, it is unclear how much he was really gesturing at a class of views and how much he was simply pointing to Nowell Smith. After all, he *twice* footnotes Nowell-Smith (1948; 1954) as a proponent of the view. Given the stingy citation practices of the era, this is notable both in its repeated specificity but also because those footnotes count for plurality of the total citations in the piece, and a majority of citations, if we discount as passing reference to Hume.

[5] Arneson calls Smart's view "hard soft determinism," and reads him as a revisionist compatibilist (2003, 233). Talbert (2016) and McKenna and Pereboom (2016), among others, read Smart as a compatibilist as well. However, the analysis that McKenna and Pereboom offer suggests one reason for reading Smart as an incompatibilist. McKenna and Pereboom claim that "Incompatibilists would not regard the control required for moral responsibility in [the sense employed by Smart and others] to be incompatible with determinism, and thus it is open to free will

something being a theory of moral responsibility. One could hold that what makes a theory about responsibility is that it does a satisfactory job of capturing our *concepts*, *thoughts*, or *beliefs* about the term 'responsibility'. (The notion of "capturing" admits of variation, and we needn't fix a specific conception. For some, capturing involves descriptive accuracy; for others, it will emphasize rational or normative appeal.) This *conceptualist* view animates many incompatibilist views, and at least some compatibilist accounts as well.[6] On a conceptualist picture of responsibility, one according to which desert talk is one of the elements essential to something being a theory of moral responsibility, Smart's positive account is no theory of responsibility at all.[7]

In contrast, a conceptualist who thinks that desert is *not* essential to responsibility (McKenna 2012), can agree that Smart's account may be read as a form of compatibilism. His rejection of desert is no barrier to capturing whatever other theoretical features figure in the demarcation of the theory being about moral responsibility. However, there is another way of thinking about the proper target of a theory of moral responsibility, according to which Smart's account may be read as a compatibilist one.

On a *phenomenalist* view about responsibility, what makes something a theory of responsibility is not its capturing of some privileged concepts, thoughts, or beliefs, but instead, its capture of the phenomena out in the world (and in our psychologies) that we take to figure prominently in our moralized praising and blaming.[8] On such accounts, what makes something a theory of moral responsibility is that it is about the psychology and social economy of blame, and in general, the

skeptics to endorse these senses" (2016, 263; Cf. Pereboom 2014, 125-138). Notice that this suggests that, by McKenna and Pereboom's reading of incompatibilism (and contrary to how they label him), Smart is more plausibly read as an *in*compatibilist. This should be unsurprising: Smart's interest in an exclusively forward-looking account of responsibility is introduced as a *consequence* of an antecedent conclusion that desert-invoking responsibility is incompatible with determinism.

[6] It is especially evident, for example, in views that hold that what makes a theory of responsibility a theory of responsibility is that it captures some (inevitably theoretically and not demonstratively-specified) notion of desert (e.g., Pereboom 2001, 2017; Caruso and Morris 2017) or self-creation (G. Strawson, 1994), properties that do not readily submit to ostension.

[7] Critical reactions to Smart often seem to presuppose an antecedent commitment to a conceptualist construal of responsibility. Watson reads the general reaction to Smart's proposal as it being "so off-centre that they in effect change the subject" (Watson 2003, 24). Similarly, Pereboom maintains that any view of responsibility that does not include desert at the center is simply not at issue in the free will debate (2017, 260). To my mind, these claims are most naturally read as expressing a conceptualist understanding of the target notion in the debate, although this reading is not incontestable.

[8] For arguments in favor of what I am here calling the phenomenalist view about responsibility, see Vargas (2004, 2015, 2017, forthcoming). This approach builds on insights from referentialist semantics, including the idea that we can isolate some reference-fixing thing—something that does not work exclusively through our thoughts—to fix reference. A proposal about that thing—e.g., the "work of the concept"—provides an independent specification of a conceptual role or functional characterization captures the stake in debates about free will and responsibility. Individual theories about responsibility are thus construed as candidates for what best fits that role or characterization.

stuff of our holding one another responsible.[9] According to the phenomenalist about responsibility, the compatibility debate is really a debate about whether our existing practices—things which are typically demonstratively specifiable—retain their normative and rational integrity if determinism is true. To the extent to which Smart captures *those* things—the phenomena out in the world that one takes to be the subject of a theory of responsibility—then Smart's account would count as a compatibilist theory of moral responsibility. However, Smart's shift to dispraise, a shift away from blame, means that his theory is no orthodox version of compatibilism according to which our practices remain largely intact. Rather, it amounts to a revisionist compatibilism that seeks to alter our practices.[10]

Although the interpretive issues surrounding Smart's account are interesting in their own right, it also prefigures a number of debates that recur in the contemporary literature. Among those issues are putative centrality of desert, the basis on which we count something as a theory of responsibility, and how we should understand what is central to debates about moral responsibility.

This section has described some of the main features of consequentialist approaches to responsibility that dominated the pre-Strawsonian Anglophone philosophical literature on moral responsibility. However, any account of this period would be remiss without a note about Strawson's "Freedom and Resentment." More than anything else, Strawson's essay undermined the credibility of consequentialist approaches to moral responsibility in the Anglophone philosophical literature. It is perhaps no small irony that a number of recent readers of Strawson's work—all members of the recent resurgence of instrumentalist approaches—have suggested that core features of Strawson's account can be accommodated within a consequentialist framework (Vargas 2008, 2013), or even that Strawson's original account, properly understood, just is a defense of a more nuanced form of consequentialism about responsibility (McGeer 2014, Miller 2014a). If the consequentialist-friendly reading of Strawson is correct, then Strawson's legacy has been, in some sense, an undoing the very kind of theory he may have set out to defend.

4. Contemporary approaches

At the start of the 21st century, consequentialist approaches were mostly regarded as non-starters. As Arneson put it, "This is the position everyone loves to hate" (2003, 233). Consequentialist theories were held to be: (1) extensionally defective, unable to distinguish between responsible agents and non-responsible agents (for example, adults, children, and dogs might all be influenceable by blame); (2) unable to distinguish between blame and other means of influencing agents; (3) products of a confusing our judgments of responsibility with the appropriateness of their expression (Scanlon 1988); (4) unable to accommodate the distinctive role of blame (Strawson 1962, Bennett 1980); (5) unable to account for backward-looking blame (Dworkin 1986); and (6) infelicitously saddled with consequentialism about normative ethics.

[9] The phenomenalist view received its canonical statement in Strawson's (1962) "Freedom and Resentment," and in his admonition to avoid "over-intellectualization" of the phenomenon, and the corresponding injunction to attend to the moral psychology and social economy of praise and blame.

[10] On Smart's relationship to revisionist approaches to free will and moral responsibility, see Vargas (2011); for more on revisionism in general, see McCormick (2016) and Vargas (forthcoming).

The rehabilitation of instrumentalist approaches to responsibility has happened in no small part because proponents of these views have tended to find those objections surmountable. Moreover, the recent trend of re-reading Strawson's own account as friendly to instrumentalist approaches (as in McGeer 2014 and Miller 2014a) has cast some of the traditional concerns about theses approaches in new light. Rather than recapitulating the back-and-forth of those historical objections—for discussions more in that vein, see Arneson (2003); Vargas (2008; 2013) and Miller (2014b)—this section focuses on five post-Strawsonian developments that have loomed large in different parts of the instrumentalist literature: the separation from consequentialist normative ethics; the idea of blame as building out or enhancing the moral or rational powers of agents; a more elaborate theory of exemptions; the development of revisionism as an explicit methodological position; and the invocation of a distinction between questions internal and external to the theory of moral responsibility.

First, because it figures prominently in what follows, it bears repeating that contemporary instrumentalist accounts have tended to emphasize that they can be (and often are) disentangled from consequentialist theories of normative ethics (e.g., Arneson 2003, 243-6; Vargas 2008; Vargas 2013a). Even so, instrumentalist accounts have benefited from making use of the considerable resources of contemporary consequentialist theorizing. For example, instrumentalist rehabilitations have appealed to both act- and rule-consequentialist formulations for the justification of responsibility practices, to indirect consequentialism, and to familiar consequentialist replies to scapegoating worries.

On Arneson's (2003) account, the appeal of responsibility practices is found in it generally being the case that holding people responsible (even in backwards-looking ways) plausibly tends to produce desirable results. These practices plausibly gain greater efficacy from us restricting our reactive responses—our *judging responses*, as he puts it—to wrongdoing of the relevant sort, by the relevant kinds of agents. According to Vargas' (2013) account, whether someone is responsible or not is settled by rules internal to what he calls "the responsibility system"—those judgments, practices, and attitudes concerned with the worthiness of moralized praise and blame. That practice can make extensive use of backward-looking assessments, and it may employ notions of desert. What makes that set of practices normatively appealing or justified, however, is that it conducive to securing the (putatively valuable) goal of fostering and refining our ability to recognize and respond to moral considerations. Victoria McGeer's (2013, 2015, 2019) formulation of instrumentalism holds that our practices of moralized praising and blaming are normatively grounded in whether so blaming enhances the wrong-doers ability to be suitably responsive to moral reasons (see also McGeer and Pettit 2015). Our capacity to recognize and respond to moral considerations is an elastic, socially-scaffolded capacity that relies on our blaming in order to build out that capacity (see also Fricker 2016, Jefferson 2019).

A second and related development has been greater attention to the way blaming practices build out the moral powers of agents through a process of "responsibilizing" (McGeer 2013), "agency cultivation" (Vargas 2013), or "prolepsis" (Fricker 2016; McGeer 2019). For many instrumentalists, a central function and justification for responsibility practices is that these practices enhance the capacity of those agents to recognize and respond to relevant moral considerations, broadly proleptic effects are unavoidably central to these accounts. Blame is proleptic inasmuch as the effect of blame leads suitable agents to have reasons or rational capacities that were either not had or not actively recognized by the agent in the moment of wrongdoing. This is the sense in which these accounts focus on cultivating a particular kind of agency though a process of "responsibilizing" practices of moral praise and blame.

However, instrumentalists differ on where to locate proleptic effects. Many instrumentalist accounts emphasize the proleptic aspect in individual instances of blaming (as in McGeer and Pettit 2015; Fricker 2016; McGeer 2019; Jefferson 2019); some emphasize its systemic effects (as in Vargas 2013, 2020). Although most contemporary instrumentalist accounts avail themselves of the idea of blame that aims to enhance moral considerations-sensitive agency, proleptic effects are plausibly more pronounced on accounts that emphasize individualized prolepsis, and less pronounced on accounts that appeal to the systemic effects of blame on groups of agents, or that otherwise limit the scope of prolepsis to a class of independently specified responsible (or non-exempt) agents.

This latter thought, about a potentially independent basis of candidates for responsibility practices, has been the subject of a third notable development in the literature. Instrumentalist accounts have taken more seriously the need to provide an account of responsible agency, that is, an account that identifies the kinds of agents that are proper candidates for our responsibility practices, attitudes, and judgments. (Alternately, one may think of this as a theory of exemptions, if we use Watson's (1987) terminology.) An articulated theory of responsible agency is important for instrumentalist approaches, in part because one traditional challenge to these accounts was that the extension of the influenceable and the extension of responsible agents was not the same. One solution is to simply invoke an independent account of responsible agency (as in Vargas 2013, which appeals to a form of reasons-responsiveness). However, as McGeer (2015) has noted, this raises concerns about the basis of that independence. A different strategy is to accept that responsible agency involves a tight connection to agency susceptible to moral influence, but to then shore up this account by specifying the nature of that susceptibility to influence—e.g., as susceptibility to the specifically reactive attitudes, or to the moral import of the expression of reactive attitudes. Concerns about extensional adequacy and the costs of revisionism on this issue remains a matter of ongoing discussion (Jefferson 2019).

Given the foregoing, it is perhaps unsurprising that a fourth development that has shaped contemporary instrumentalist accounts is the gradual emergence of an explicitly revisionist approach to moral responsibility. On one standard definition of revisionism about moral responsibility, a theory of moral responsibility is revisionist if the truth of the theory's account of moral responsibility is in conflict with commonsense views about that thing.[11] The appeal of revisionism for instrumentalists should be apparent: if one has a principled basis for revisionism, it may offer a principled solution to lingering gaps between the extension of ordinary judgments about moral responsibility and the instrumentalist's account of responsibility.

[11] Compare McCormick's formulation of revisionism: "Revisionism is the view that we can and should distinguish between what we think about moral responsibility and what we ought to think about it, that the former is in some important sense implausible and conflicts with the latter, and so we should revise our concept accordingly" (2016, 109). For overviews of revisionism, see McCormick (2016), McKenna and Pereboom (2016, pp. 286-293); and Vargas forthcoming.

The core of this idea was clearly present in Smart's (1961) account, and it has been picked up and developed in the current crop of instrumentalist theories. For example, Arneson claims that instrumentalism "has to be regarded as a substitute for the ordinary idea of responsibility . . . A precondition of finding influenceability acceptable is having good grounds for finding the ordinary notion of responsibility unsustainable. So the fact that influenceability does not mesh perfectly with the ordinary notion of responsibility is not *per se* an objection to it" (2003, 249). Versions of revisionism about free will and/or moral responsibility can also be found in the instrumentalist accounts endorsed by Vargas (2013), McGeer (2015), and Jefferson (2019). Of course, one need not be a revisionist to be an instrumentalist, and there are revisionists who are not obviously instrumentalists (Nichols 2015; Doris 2015a).

Smart, Arneson, and Vargas are all explicit that their revisionism is motivated by the thought that folk thinking about responsibility is at least partly committed to an implausible libertarian metaphysics of agency. One need not be a revisionist to be an instrumentalist, though. One might think that folk understandings of responsibility have always been compatibilist (this is perhaps the traditional approach to compatibilism—see Vargas 2011). On that account, instrumentalism just is an account of the normative foundations of the practices as we have them.

Partly connected to the question of whether various instrumentalist accounts are revisionist, there is the matter of what is to be revised. Vargas' indirect, two-tiered account is intended to limit or avoid the need for notable revisions in our ordinary practices. The revisions in his account are primarily at the level of the kinds of things we *think* and *believe* about the basis of our responsibility practices. In contrast, more direct forms of instrumentalism tend to more readily accept the need for important revisions in our practices, that is, in what we *do* (as in Arneson 2003; McGeer 2013; Jefferson 2019, and if we count his positive proposal as revisionist, especially so in the case of Pereboom's 2014 proposal).

A fifth and final innovation worth noting is that some instrumentalists have employed a distinction between the justification of responsibility norms and their content. Something like this distinction is suggested in "Freedom and Resentment," in Strawson's distinction between questions internal and external to the responsibility practice. Building on later remarks by Strawson (1985), Dale Miller (2014a) has argued that the idea that instrumental concerns are the wrong kind of answer to questions about responsibility was intended by Strawson as an invocation of an idea in Carnap, namely, the idea that there are some questions whose answers are internal to a framework, and that it is a separate question to ask about the appeal of having that framework.

The internal/external to the framework idea seemed to be in the air at the time, with a related idea figuring in the work of Rawls (1955) and HLA Hart (1959). On their accounts, we ought to distinguish between the justification of rules and institutions (which may be given in instrumental terms) and the content of the rules or the norms of the institutions (which may be backward-looking, desert-based, or otherwise non-instrumental in character). Vargas (2013; 2015), in particular, has made use of this distinction in distinguishing between the instrumental character of the responsibility system, and the potentially desert-based, backward-looking character of first-order (or substantive) responsibility norms.

This section canvassed five ideas that have figured prominently in contemporary instrumentalist approaches— the independence of instrumentalism from consequentialism; the agency cultivating or proleptic aspect of responsibility practices; more attention to responsible agency; the evolution of revisionist theorizing; and arguably greater clarity about questions internal and external to frameworks. As we will see in the next section, these ideas tend to figure in how contemporary instrumentalist accounts respond to various standard challenges.

5. Objections and debates

In this section, I discuss several standard objections and ongoing debates for instrumentalist approaches. These include the question of scapegoating, objections that instrumentalists are relying on the wrong kind of reason, concerns about self-effacement, and the objection that instrumentalists are unable to capture the notion of desert that figures in debates about free will and moral responsibility.

5.1 Scapegoating

One traditional complaint about consequentialist theories of ethics is that they seem to support scapegoating, or the punishing of people who are innocent. One might wonder whether instrumentalist approaches to responsibility are vulnerable to a parallel complaint.

Whatever the prospects are for consequentialist theories of normative ethics, instrumentalist accounts have good resources for deflecting concerns about the equivalent of scapegoating in a theory of moral responsibility—in this case, blaming the non-responsible when it is expedient to do so. Two replies, at least, are available: the *definitional* response and the *modularity* response. They can be employed jointly or individually, depending on the particulars of the account.

The *definitional* response to scapegoating is well-expressed by Arneson: “one is responsible for an act if one did it and doing of this sort are influenceable by blaming or punishing. One cannot squeeze hard on this admittedly thin notion of responsibility to somehow induce it to imply that one can be responsible for a crime one did not commit, because one’s doing it is by definition required for responsibility” (2003, 245). One might put pressure on Arneson’s formulation in various ways. For example, depending on how one thinks of an agent’s doings, omissions could be a problem for the idea that one’s doing is required for responsibility. Still, the underlying point is plausible: whatever one thinks responsibility is, it requires that some outcome be suitably related to the agent’s behavior. Adding to that thought the *further* thought that the agent must able to be influenced by blame, or that it must be the case that rules licensing blame must generally produce better responsiveness to moral considerations, does not eliminate that conceptual or definitional point.

Of course, one can ask why *that* conceptual constraint? To this, the instrumentalist might reply that this new question is no longer a question about when someone is responsible. Rather, it is a question about why we should have a notion of responsibility at all. Perhaps we do better to give up responsibility and go in for some other thing, something that does an even better job at achieving the instrumentalist’s end. However, if this is the question, it is no longer the instrumentalist who can be accused of changing the subject. Thus, the instrumentalist may demur, insisting that her account is merely an account of responsibility, and that it is beyond her ambition to offer an account of what collection of practices, all things considered, best serves the identified instrumental end. I leave it to others to decide whether this is a satisfactory response.

Whatever the answer might be, there are further things that the instrumentalist can say about these issues. In response to the "why responsibility and why not something else?" question, instrumentalist can appeal to some of the same considerations suggested by Strawson in "Freedom and Resentment." It might be that, for example, given our psychologies, no nearby arrangement of practices would do. Alternately, it might simply be a matter of feasibility—perhaps there are normatively more appealing arrangements, but they are not readily available to us. If so, the most we can effectively do is to militate for modest adjustments of the practices we find. Or, one might be moved epistemic concerns, citing concerns about unintended consequences in transforming practices in some more radical way. This might seem especially plausible if one was inclined to think that the underlying affective psychology was ineradicably entangled with other things—love, human affection, social coordination, and so on. If so, the question of the comparative appeal of this set of practices [including responsibility and its entanglements] and that set [devoid of responsibility along with transformations in the things with which it is entangled] cannot be readily answered. One might even hold, as Strawson seems to have, that the only answers that can be given are always internal to some or another set of value-fixing affective orientations.

The point of the foregoing remarks is not to take a stand on how that conversation should go. It is simply to note that it is not obvious that the instrumentalist faces a unique or distinctive challenge in employing the definitional response to the scapegoating objection. If the instrumentalist elects to take up subsequent "metanormative" questions about those conceptual constraints, there are widely utilized ideas to which she might appeal to explain why we should embrace responsibility, rather than joining eliminativist in calling for its ejection from our moral lives.[12]

The *modularity* response was already highlighted in §2, without the benefit of a label. The idea is as follows. A theory of moral responsibility is a theory of moral responsibility, and not some other thing. To be an instrumentalist about moral responsibility need not entail that one is a consequentialist about normative ethics. Some instrumentalist accounts forge a relatively tight link between susceptibility to the influence of moral blame and being blameworthy (e.g., Miller 2014a; McGeer 2019; Jefferson 2019). Others do not (Vargas 2013). However strong that link, the *normative* appeal of scapegoating—blaming those who are not blameworthy; or even, expanding the scope of the putatively blameworthy to those whom it is simply expedient to blame—depends on one's wider philosophical commitments.

If one accepts consequentialism, then it was already normatively appealing to think that when the payoffs are high enough, it may do to blame those who are not blameworthy (or in general, to expand the class of the blameworthy to the expediently blamed). Similarly, if one's normative

[12] Here, the issue of revisionism lurches into the debate again. One may hold that there are some benefits for moralized praise and blame, so long as they are denuded of any commitments to, say, libertarian conceptions of free will. Such views have typically been taken up under the banner of offering explicitly revisionist theories of moral responsibility which aim to (sometimes selectively) purge commitments that involve incompatibilism, implausible psychologies, or overly-demanding commitments on desert within our conceptual or practical lives (e.g., Smart 1962; Hurley 2003; Vargas 2013a; McCormick 2015; Nichols 2015; Doris 2015a). As I understand it, the positive account in Pereboom (2014), but perhaps not Pereboom (2001), is best construed as an instance of this approach—albeit one distinguished by its eschewal of desert. More about this below.

ethics funds a prohibition against blaming those for whom it is merely expedient to do so—suppose that agents should always be treated as ends in themselves, and not merely as means—, then one already has the resources for explaining why scapegoating the non-responsible is impermissible. These prohibitions do not disappear because one's theory of moral responsibility is justified in instrumentalist terms. Instead, the normative ethical theory can be understood as providing side-constraints on the scope of the instrumental reasons.

5.2 The wrong kind of reasons

In a famous passage in "Freedom and Resentment," Strawson imagines the instrumentalist's interlocutor saying "the only reason you have given for the practices of moral condemnation and punishment in cases where this freedom is present is the efficacy of these practices in regulating behavior in socially desirable ways. But this is not a sufficient basis, it is not even the right *sort* of basis, for these practices as we understand them" (1962, 4).

This observation is sometimes taken as the *locus classicus* of the "wrong kind of reasons" objection against instrumentalist theories. It prominently figures in Darwall's (2006) claim that "*Desirability is a reason of the wrong kind to warrant the attitudes and actions in which holding someone responsible consists in their own terms*" (15). "Strawson's Point," as Darwall calls it, is supposed to identify a fatal problem for instrumentalist theories.[13]

It bears noting that immediately after introducing the pessimist's accusation that the optimist is relying on the wrong kind of reason, Strawson makes it clear that *his project is to give the optimist more to say*. On the face of it, Strawson did not seem to think the objection was fatal to instrumentalism (Cf. McGeer 2014, Miller 2014a). Looking at the details of Strawson's account suggests one reason for thinking that the wrong reasons objection can be met by the instrumentalist.

Recall the idea, noted above, that Strawson had a roughly Carnapian picture of responsibility practices. There are questions *internal* to the practice, and there are questions *external* to the practice. On this reading, what makes instrumental considerations the wrong kind of reason is that they raise questions external to the practice. For Strawson, the idea that social efficacy is not the right basis for the practice is a point internal to the practice. If instrumental reasons have a place, it is external to the practice. This is, of course, exactly what at least some contemporary instrumentalist maintain (Vargas 2008, 2013, 2015; Miller 2014a).

Here is what the "wrong kind of reasons" objection plausibly gets right: consequences are mostly irrelevant to the propriety conditions of blame, gratitude, and resentment. Each of these things has its own aptness or propriety conditions, conditions that don't appeal to consequences. Consequences may speak to whether it is *advantageous* to express our blame, gratitude, or resentment, but not to whether those attitudes are *merited*, *deserved*, or *proper*. These things are, for Strawson, in large part structured by a relatively fixed set of reactive dispositions that are keyed to perceptions of whether others have exercised due concern or adequate quality of will.

[13] For a helpful discussion of why one might doubt that Darwall's account has adequate resources for motivating a wrong kind of reasons objection against instrumentalists, independent of Strawsonian concerns, see Miller 2014a. Even if we accept Darwall's account of the closed "second-personal" circle of accountability, the considerations in the present section suggest reasons that at least some instrumentalists could accept this constraint.

One can accept all of the preceding, while allowing that instrumental questions might arise at a different order of inquiry. Many norm-structured practices have exactly the structure to them. Foul calls in a sport are typically justified by the thought that the safety of the players must be preserved, but this must be balanced with the enjoyment of spectators and players in the flow of the game. However, whether a particular play is a foul or not is clearly not settled by appeal to those framework questions. They are settled internal to the framework, by appeal to the rules of the game. In the context of the game—within the framework of participants, one might say—questions of safety and enjoyment are irrelevant to whether something is a foul (unless the rule specifies that those are considerations). We can, of course, step outside those rules and ask if we want to have those rules, whether there is a better set of rules available to us, and whether we have reason to allow ourselves to be bound by those rules. But the rules are the rules, at least until we change them.

Can the instrumentalist exploit this idea to make sense of the role of instrumental considerations? Strawson entertains the idea that we might step outside the "reactive" or participant framework, and take up a standpoint of objectivity, but he was cautious about the possibility of undertaking normative inquiries external to the participant stance.[14] The suggestion in "Freedom and Resentment" is that it is unclear where we might stand we take up framework questions, and on what basis we might adopt a different set of attitudes and practices.

Here, though, instrumentalists may depart from Strawson in his holding that, for example, normative questions bottom out in our attitudes, or that the attitudes that ground normative matters rise and fall together, or that one cannot privilege some of these attitudes, and not others, or that one might identify some axiological notions as the basis on which to evaluate various attitudes and practices. An instrumentalist who appeals to non-Strawsonian views on any of these matters might find grounds to insist that there is some place we can stand outside the practice of responsibility to assess whether we might retain it, reject it, or revise it.

What the proponent of the wrong kind of reasons objection needs is for it to be impossible to intelligibly take up questions about normative frameworks that are external to that framework, or at least, some special reason to think one cannot do so in the context of moral responsibility. That many of our normative practices permit these sorts of questions, and that at least some instrumentalist accounts mimic that structure, suggests that the wrong kind of reasons objection is, so far, the wrong kind of objection to make against indirect or two-tiered instrumentalist theories of moral responsibility.

The existing wrong kinds of reason objection is most promising against direct and classical consequentialist accounts of the sort discussed in § 3, above. Instrumentalisms that recognize a difference between questions internal and external to a framework, or that can distinguish between the contents and the justifications of practices and institutions (Cf. Rawls 1955; Hart 1959) have resources for deflecting the concern.

5.3 Self-effacement

The foregoing remarks can give rise to concerns about self-effacement (Doris 2015b; Miller 2014b). There are different ways to put the concern, but in the abstract the idea is that

[14] In general, the question of normativity in Strawson is an elusive one (cf. Watson 1987).

instrumentalism, particularly those that allow for indirectness at the level of first-order practices, can produce an unacceptable bifurcation in the thinking of blamers. The worry is that blamers must employ a set of first-order principles about blame, while simultaneously recognizing that there is a second-order set of concerns about whether these principles ought to be amended or otherwise adjusted. This concern has taken various forms, including skepticism about the psychological tenability of accepting two orders of concern and the risk that this sort of "divided" thinking might de-motivate the efficacy of the first-order commitments (Williams 1988).

The details vary by instrumentalist theory, but one view to which instrumentalist can avail themselves is the idea that first-order principles, beliefs, and judgments may be most effective when they are intuitive, routinized, and habitual (Arneson 2003). Routinized or habituated principles will tend to shape one's dispositions for perceiving what is morally significant (Vargas 2013, Miller 2014b, McGeer 2019).

If the locution can be forgiven, we can begin by noting that the Williams-style worry about self-effacement has two faces. There is the point about the psychological tenability of two level thinking, and a worry about its motivational efficacy. Experience speaks to its tenability. As John Doris has observed, it does not seem impossible for one to think that one shoots a three pointer and to also hold "the justification of participating in a practice where things are done for the sake of winning games is that this participation ethically improves the participant" (2015b, 2632). On the matter of motivational efficacy, the issue is that awareness of the possibility of second-order concerns will impair the ordinary efficacy of otherwise entrenched habits of mind. This worry is especially acute for act-consequentialist style versions of instrumentalism where the propriety of blaming in any instance depends on its effects.

First, it is not obvious that instrumentalists must require that agents be aware of the distinction between the first order norms of responsibility-holding and the normative foundations of that practice. Further, is simply unclear that much follows from awareness of this distinction. Whether this awareness would in any interesting way alter the character of the first-order norms (or the typical person's relationship to them, for that matter) seems to be an entirely empirical question.

Second, Dale Miller (2014b) has argued that Strawson's account of the reactive attitudes provides resources for instrumentalists in this context. To the extent to which instrumentalist accounts are relying on first-order judgments and principles that correspond with the alignment of our affects—a matter about which many contemporary instrumentalists are explicitly committed—then the frame of mind of the instrumentalist is one where the reactive attitudes are governing in their usual way. If there is any special demand on the mind of the instrumentalist, it may be as minimal as resolving to inhabit those attitudes, and to let them fully become habits of mind, subject to disavowal only under isolated conditions. As Strawson himself emphasizes, we can step back from our attitudes and ask questions about them, and whether they might admit of alteration. That is, however, a different frame of mind than the frame of mind where one is operating within the space of the attitudes as they present themselves to us. Of course, radical departures from the dispositions of our reactive attitudes may be difficult to sustain, but the pressure towards conservatism about revision tends to be recognized by many contemporary instrumentalists.

5.4 Desert

What about desert? Some philosophers have allowed that there may be multiple senses of responsibility, but that the one that figures in debates about free will and moral responsibility concerns some or another version of the idea of desert (Pereboom 2001, 2014, 2017; Caruso and Morris 2017). However, desert is a notion with no central role in consequentialist theories of ethics. So, one might think, it is difficult to see how desert might operate in instrumentalist theories that emphasize the consequences of the practice in establishing their justification. Pereboom has argued that instrumentalists of various stripes do not offer basic desert theories of responsibility (2014, 2; 2017, 260). Perhaps this is true of some instrumentalists (e.g., Dennett 1984; Arneson 2003), but at least some instrumentalist theories have resources for capturing the operative sense(s) of desert.

According to Pereboom's justly influential account, the notion of desert that is at stake in debates about free will is something he labels *basic desert*, where the operative notion of desert is one according to which an agent "would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations" (2014, 2).

The issues here are delicate, but the general contours of the main options for the instrumentalist are relatively clear. First, it is open to instrumentalists (and others) to reject the idea that desert is essential to moral responsibility.[15] Second, instrumentalist might accept the basic desert conception, and argue that it can be met. Third, instrumentalist might accept that some notion of desert is involved in attributions of responsibility, but not the one that figures in Pereboom's characterization of basic desert.

In what follows, I'll focus on the second and third approaches, as they have figured in most contemporary instrumentalist accounts. Few instrumentalist accounts have explicitly repudiated any notion of desert.[16]

Some accounts have maintained that the basic desert condition can be met by at least some instrumentalist accounts (Vargas 2015). To see why, we must first get clear on the dialectical burdens in this context.

On the standard construal of basic desert, the idea is that desert is generated by the nature of the agent and the action, and blaming isn't licensed by consequentialist or contractualist considerations. This characterization cannot, by itself, settle the question of whether responsibility is compatible with determinism or not. If basic desert is, from the outset, construed as an incompatibilist notion, it is a non-starter in its primary dialectical function, namely, as a characterization of the notion of responsibility in the dispute. Of course, Pereboom and other

[15] McKenna (2012), who is not an instrumentalist in the sense under discussion, has explored a version of this view/ He ultimately argues that a desert-invoking formulation is compatible with his account of responsibility. David Shoemaker (2015) has suggested that questions of desert are restricted to harsh treatment and not central to responsibility.

[16] In recent work, Nelkin (2016) has argued that although 'desert' has been used in variety of ways, desert and accountability are mutually entailing. If this is right, then quite apart from whether we accept Pereboom's specific account of basic desert, it may not be open to proponents of a theory of accountability to insist that desert has no role to play.

putative incompatibilists think that the only way to secure basic desert, and thus moral responsibility in the relevant sense, is via some form of agency incompatible with determinism. Similarly, compatibilists who accept the basic desert characterization must think that their accounts can produce basic desert. The point here is that there is a further debate to be had, and that basic desert's supposed entailment of incompatibilism cannot be definitional, on pain of basic desert failing as a neutral characterization of the stakes of the debate.

So, can instrumentalists offer an account of responsibility that, on the face of it, meets the conditions of basic desert? Compatibilist who are not instrumentalists in the sense at stake here have offered such accounts. For example, David Brink maintains that desert is to be understood as the product of two independent variables, wrongdoing and culpability (2012, 498; see also Brink and Nelkin 2013).[17] The issue is whether instrumentalists can help themselves to this sort of account, or something like it.

Here, one's variety of instrumentalism seems to matter. For instrumentalist accounts that entirely ground desert talk in consequences, then Pereboom is right that such accounts fail to be basic desert accounts, because to deserve praise or blame would be determined "merely by virtue of consequentialist or contractualist considerations." However, two-tier versions of instrumentalism seem not to run afoul of that prohibition.

Recall that on a two-tiered instrumentalist account, the norms, judgments, and attitudes that make up the responsibility system can be backward-looking. On this sort of account, the propriety conditions for deserving blame can be the moral qualities of the agent and the act. Thus, the fact that the agent deserves blame is *not* merely by virtue of consequentialist considerations. Indeed, the two-tiered theorist can allow that *no* judgments of blame are settled "merely by virtue of consequentialist considerations." Instead, they are settled by, for example, whether the agent has acted wrongfully, and whether the agent was culpable in so acting. The incompatibilist might contest the instrumentalist's construal of culpability and acting. Even so, the constraint that the deserving of blame can be established independent of its consequences can readily be satisfied by a two-tiered instrumentalist theory.

Here's what's the basic desert-ers "no consequentialism" constraint rightly captures: there is a conception of desert—one important to many of our social and more specifically moral practices—according to which consequences are irrelevant. Trying to capture that notion of desert with a conception of desert that is sensitive to consequences is bound to do considerable violence to this ordinary and central notion of desert (Doris 2015b). Inasmuch as a consequence-insensitive notion of desert is implicated in attributions of moral responsibility, then at least one variety of instrumentalism about moral responsibility can capture this notion of desert. In doing so, the two-tiered instrumentalist captures something of the flavor of Strawson's observation that there are questions internal to a practice, and questions external to a practice. Questions about basic desert are questions internal to the practices of responsibility. On the two-tiered version of

[17] One further virtue of Brink's account is that it also provides a compatibilist-friendly way to capture the core retributivist idea, that punishment is proportional to that desert. So, incompatibilists convinced that the stakes of the responsibility debate must also involve a notion of desert or responsibility that supports retributivism (cf. Caruso and Morris 2017) cannot readily dismiss this version of compatibilism as failing to even join the putative debate. It is notable, however, that not all compatibilists have thought it a desideratum that any notion of desert captured by their account can support retributive practices, and practices of punishment (e.g., Scanlon 1998 and Wallace 1994).

instrumentalism, instrumentalism is constrained to the grounds we have for that practice as a whole, and so its internal-to-the-practice account of desert can be compatible with basic desert.

Are there other ways a (potentially non-two-tiered) instrumentalist might accept the demand for a basic desert notion of responsibility? Following the definitional strategy suggested by Arneson's (2003) remarks about scapegoating, an instrumentalist could accept that basic desert is a conceptual requirement on responsibility, and that it operates on non-consequentialist terms. So, the constraint of basic desert would be accepted on conceptual grounds, settled by the definitional features of responsibility, and not because of consequentialist considerations. It would then be open to the instrumentalist to then add that the reason for keeping this package of conceptual commitments is that it produces instrumentally valuable results. This wouldn't require endorsement of a two-tiered instrumentalism, but it would raise some of the same questions that occur in the context of scapegoating, e.g., raising questions about why we should accept basic desert practices vs. some other practice that does without them.[18]

It is unclear how far apart these strategies are from one another. What matters for present purposes is this: on the face of it, instrumentalist accounts of moral responsibility can satisfy basic desert construals of the philosophical stakes. Whether the resultant conception of basic desert has other unappealing features, or whether one accepts the implied account of culpability is a further matter.

As noted at the outset of this subsection, there are other routes available to the instrumentalist seeking to address concerns about blame. Instrumentalists can, for example, accept that desert figures in responsibility attributions, but reject the claim that the stakes are *basic* desert. Would such a position mean that one is willfully ignoring the central philosophical dispute about moral responsibility? It need not. Recall the difference between conceptualist and phenomenalist construals of the philosophical stakes of a theory of responsibility. Many proponents of the importance of basic desert are most naturally interpreted as conceptualists, that is, endeavoring to capture our *concepts*, *thoughts*, or *beliefs* about the term 'responsibility'. However, the instrumentalist can avail herself of the phenomenalist idea that this is a methodological mistake. Instead, on the phenomenalist reading, the proper philosophical stakes are about the nature and normative integrity of our practices.[19] These practices, she might say while gesturing at various phenomena in our social world—and not some armchair stipulation of a metaphysics of desert—are the subject of my account, and plausibly the subject of most accounts in the long history of philosophical debates about responsibility. For the phenomenalist, if it turns out that what we

[18] Pereboom's desert-free account of responsibility (2014) might be understood as a contribution to this debate.

[19] It is an error, I think, to reply that accounts of this sort make compatibilism too easy, or that they eliminate any substantive difference between compatibilism and hard incompatibilism (this thought is in the spirit of remarks in Pereboom 2014, 2-3; 2017, 260). Nothing on phenomenalist construals of the debate rule out the possibility that our ordinary practices may presume that we have impossible or unlikely forms of agency, or that our practices are normatively indefensible. The former is claimed by some incompatibilists and many revisionists and disputed by conventional compatibilists. Disagreements about latter is what separates typical revisionists and conventional compatibilists from eliminativists or hard incompatibilists.

need to explain the nature, function, and normative integrity of our practices as we find them is some non-basic notion of desert, then so much the worse for basic desert.[20]

A different but complementary strategy is to lean on the revisionist impulses that oftentimes accompany instrumentalist accounts. On this approach, one could grant that basic desert captures the ordinary understanding of desert. However, one can also maintain that some attenuated notion of desert is the proper successor notion (as in Smart 1963 and Arneson 2003) or that we can make do with forms of responsibility that do away with desert (as in Pereboom 2014;).

Although there are a number of paths open to instrumentalists in addressing concerns about basic desert, it is also true that one might place a special value in securing even more demanding forms of desert, including pre-institutional or practice-independent forms of desert, or desert of the "making sense of heaven and hell" variety (Strawson 1994). Whatever the attractions of these other forms of desert may be, they are less plausible as neutral characterizations of the stakes of the responsibility debates. They are more plausible as substantive accounts of more particular conceptions of desert, for which even the evaluative criteria for these notions—e.g., normative appeal, folk recognizability, conceptual importance, etc.—remains a matter of robust theoretical dispute.

Conclusion

Instrumentalist theories of moral responsibility are enjoying something of a renaissance. Given that these accounts have resources for addressing the traditional worries raised against them, and given their recent proliferation in the literature, there is some reason to think that these accounts are (rightly or wrongly) once again an important approach for understanding the nature of moral responsibility.

[20] I take it that something like this is the spirit of how McGeer and Funk (2017) think we should understand the significance of findings that show that our putatively retributive attitudes are sensitive to consequences.

Bibliography

Arneson, R. J. (2003). The Smart Theory of Moral Responsibility and Desert. In S. Olsaretti (Ed.), *Desert and Justice* (pp. 233-258). Oxford: Oxford.

Bennett, J. (1980). Accountability. In Z. Van Straaten (Ed.), *Philosophical Subjects*. New York: Clarendon.

Brink, D. O. (2012). Retributivism and Legal Moralism. *Ratio Juris*, *25*(4), 496-512.

Brink, D. O., & Nelkin, D. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility*, *1*, 284-314.

Caruso, G. D. (2012). *Free will and Consciousness: A Determinist Account of the Illusion of Free Will.* Lanham, Md.: Lexington Books.

Caruso, G. (2015). Free Will Eliminativism: Reference, Error, and Phenomenology. *Philosophical Studies*, *172*(10), 2823-2833.

Caruso, G. D., & Morris, S. G. (2017). Compatibilism and Retributive Desert Moral Responsibility: On What is of Central Philosophical and Practical Importance. *Erkenntnis*, *82*(4), 837-855.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy*, *76*, 256-282.

Darwall, S. L. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Mass.: Harvard University Press.

Dennett, D. (1984). *Elbow Room*. Cambridge: MIT.

Doris, J. (2015a). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Doris, J. (2015b). Doing Without (Arguing about) Desert. *Philosophical Studies*, *172*(10), 2625-2634.

Double, R. (1996). *Metaphilosophy and free will*. New York: Oxford University Press.

Dworkin, G. (1986). Review of Elbow Room. *Ethics*, *96*(2), 423-425.

Dworkin, G. (Ed.). (1970). *Determinism, Free Will, and Moral Responsibility*. Englewood Cliffs, New Jersey: Prentice-Hall.

Fricker, M. (2016). What's the Point of Blame? A Paradigm Based Explanation. *Nous*, *50*(1), 165-183.

Jefferson, A. (2019). Instrumentalism about Responsibility Revisited. *The Philosophical Quarterly*, *69*(276), 555-573.

Hart, H. L. A. (1959). Prolegomena to the principles of punishment. *Proceedings of the Aristotelian Society New Series*, *60*, 1-26.

Hook, S. (Ed.). (1958). *Determinism and Freedom in the Age of Modern Science*. New York: New York University Press.

Hurley, S. L. (2003). *Justice, Luck, and Knowledge*. Cambridge, MA: Harvard University Press.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.

McCormick, K. (2015). Companions in Innocence: Defending a New Methodological Assumption About Moral Responsibility. *Philosophical Studies*, *172*(2), 515-533.

McCormick, K. (2016). Revisionism. In K. Timpe, M. Griffith, & N. Levy (Eds.), *Routledge Companion to Free Will* (pp. 109-120). New York: Routledge.

McGeer, V. (2013). Civilizing Blame. In J. D. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 162-188). Oxford: Oxford University Press.

McGeer, V. (2014). P.F. Strawson's Consequentialism. In D. Shoemaker & N. A. Tognazzini (Eds.), *Oxford Studies in Agency and Responsibility Volume 2* (pp. 64-92). New York: Oxford University Press.

McGeer, V. (2015). Building a Better Theory of Responsibility. *Philosophical Studies*, *172*(10), 2635-2649.

McGeer, V. (2019). Scaffolding Agency: A Proleptic Account of the Reactive Attitudes. *European Journal of Philosophy*, *27*(2), 301-323.

McGeer, V., & Funk, F. (2017). Are 'Optimistic' Theories of Criminal Justice Psychologically Feasible? The Probative Case of Civic Republicanism. *Criminal Law and Philosophy*, *11*(3), 523-544.

McGeer, V., & Pettit, P. (2015). The Hard Problem of Responsibility. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility, Vol. 3* (pp. 160-188). Oxford: Oxford University Press.

McKenna, M. (2012). *Conversation and Responsibility*. New York: Oxford University Press.

McKenna, M., & Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. New York: Routledge.

Miller, D. E. (2014a). "Freedom and Resentment" and Consequentialism: Why 'Strawson's Point" Is Not Strawson's Point. *Journal of Ethics and Social Philosophy*, *8*(2), 1-22.

Miller, D. E. (2014b). Reactive Attitudes and the Hare-Williams Debate: Towards a New Consequentialist Moral Psychology. *The Philosophical Quarterly*, *64*(254), 39-59.

Moore, G. E. (1903). *Principia Ethica*. Cambridge, U.K.: Cambridge University Press.

Nelkin, D. (2016). Accountability and Desert. *Journal of Ethics*, *20*(1-3), 173-189.

Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. New York: Oxford University Press.

Nowell-Smith, P. (1948). Free Will and Moral Responsibility. *Mind*, *57*(225), 45-61.

Nowell-Smith, P. (1954). Determinists and Libertarians. *Mind*, *63*(251), 317-337.

Nozick, R. (1981). *Philosophical Explanations*. Oxford: Clarendon Press.

Pears, D. F. (Ed.). (1965). *Freedom and the Will*. London, U.K.: Macmillan & Co.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. New York: Oxford University Press.

Pereboom, D. (2017). Response to Daniel Dennett on Free Will Skepticism. *Rivista Internazionale di Filosofia e Psicologia*, *8*(3), 259-265.

Rawls, J. (1955). Two Concepts of Rules. *Philosophical Review*, *64*, 3-32.

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Scanlon, T. M. (1988). The Significance of Choice. In S. M. McMurrin (Ed.), *The Tanner Lectures on Human Values* (pp. 150-216). Cambridge, UK: Cambridge University Press.

Schlick, M. (1939). When Is A Man Responsible? (D. Rynin, Trans.). In *The Problems of Ethics* (pp. 143-158). New York: Prentice Hall.

Shoemaker, D. (2015). *Responsibility from the Margins*. New York: Oxford University Press.

Shoemaker, D., & Vargas, M. (forthcoming). Moral Torch Fishing: A Signaling Theory of Blame. *Nous*. Retrieved from https://doi.org/10.1111/nous.12316

Smith, A. (2013). Moral Blame and Moral Protest. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 27-48). New York: Oxford University Press.

Smart, J. J. C. (1961). Free Will, Praise, and Blame. *Mind*, *70*, 291-306.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies*, *75*, 5-24.

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *XLVIII*, 1-25.

Strawson, P. F. (1985). *Skepticism and naturalism: some varieties*. New York: Columbia University Press.

Talbert, M. (2016). *Moral Responsibility: An Introduction*. Malden, MA: Polity Press.

Vargas, M. (2004). Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly*, *85*(2), 218-241.

Vargas, M. (2008). Moral Influence, Moral Responsibility. In N. Trakakis & D. Cohen (Eds.), *Essays on Free Will and Moral Responsibility* (pp. 90-122). Newcastle, UK: Cambridge Scholars Press.

Vargas, M. (2011). The Revisionist Turn: Reflection on the Recent History of Work on Free Will. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New Waves in the Philosophy of Action* (pp. 143-172). New York: Palgrave Macmillan.

Vargas, M. (2013a). *Building Better Beings: A Theory of Moral Responsibility*. Oxford, U.K.: Oxford University Press.

Vargas, M. (2013b). If Free Will Does Not Exist, Then Neither Does Water. In G. Caruso (Ed.), *Exploring the Illusion of Free Will and Moral Responsibility* (pp. 177-202). Lanham, Maryland: Lexington Books.

Vargas, M. (2015). Desert, Responsibility, and Justification: Reply to Doris, McGeer, and Robinson. *Philosophical Studies*, *172*(10), 2659-2678.

Vargas, M. (2017). Contested Terms and Philosophical Debates. *Philosophical Studies*, *174*(10), 2499-2510.

Vargas, M. (2020). Negligence and Social Self-Governance. In A. R. Mele (Ed.), *Surrounding Self-Control* (pp. 400-420). New York: Oxford University Press.

Vargas, M. (forthcoming). Revisionism. In J. Campbell, K. M. Mickelson, & V. A. White (Eds.), *The Wiley Companion to Free Will*. Oxford: Wiley-Blackwell.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (1987). Responsibility and the Limits of Evil. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions* (pp. 256-286). New York: Cambridge.

Watson, G. (2003). Introduction. In G. Watson (Ed.), *Free Will* (2nd ed., pp. 1-25). Oxford: Oxford.

Williams, B. (1988). The Structure of Hare's Theory. In D. Seanor & N. Fotion (Eds.), *Hare and Critics: Essays on Moral Thinking* (pp. 185-196). Oxford, U.K.: Clarendon Press.